Visualizing Temporal Dynamics at the Genomic and Metabolic Level

Romain Bourqui and Michel A. Westenberg
Eindhoven University of Technology, The Netherlands
{r.bourqui, m.a.westenberg}@tue.nl

Abstract

We present an application for integrated visualization of gene expression data from time series experiments in gene regulation networks and metabolic networks. Such integration is necessary, since it provides the link between the measurements at the transcriptional level and the observable characteristics of an organism at the functional level. Our application can (i) visualize the data from time series experiments in the context of a regulatory network and a metabolic network; (ii) identify and visualize active regulatory subnetworks from the gene expression data; (iii) perform a statistical test to identify and subsequently visualize affected metabolic subnetworks. Initial results show that our integrated approach speeds up data analysis, and that it can reproduce results of a traditional approach that involves many manual and time-consuming steps.

1 Introduction

Biologists face the difficult task of relating experimental data to biological processes that take place at different levels of organization. Improvements in acquisition techniques have made this task even more challenging. For example, microarrays are increasingly used to study dynamic behavior of cellular processes by capturing multiple gene expression profiles at discrete time points. The problem now is to understand the experimental data, relate it to the multiple levels of organization, verify existing knowledge, and make new discoveries.

In this paper, we focus on two particular levels of cellular organization: gene regulation networks and metabolic networks. In a *gene regulatory network*, the elements are the genes of the organism. Two genes are linked if the gene product of one of these genes regulates the other gene. Metabolism is the set of biochemical reactions that are used to perform vital biological functions such as energy generation. Each metabolic function is modelled by a set of interconnected biochemical reactions corresponding to a small graph called a metabolic pathway [19]. Since the output of a pathway is often the input of another pathway it is possible to merge all these pathways into a single network, called a *metabolic network* [11, 12].

There exist many tools in bioinformatics for visualization of these types of biological networks, see Saraiya et al. [16] and Suderman & Hallett [18] for recent overviews. Prominent tools are Cytoscape [17], VisANT [7], BiologicalNetworks [1], and VANTED [10], for general biological network visualization. Some other tools specialize on pathways [2,5,9,13,15,20,21,24]. Visualization of gene expression data from time series in these tools has been mostly limited to displaying either the whole time series in a node or coloring a node based on a single time point. Since no information is mapped to the edges, it is very difficult to understand which substructures of the network are active at any point in time. None of these tools map statistical properties of the expression data to a meaningful visual attribute. Statistical properties are very important in microarray data analysis, because they provide a measure of confidence for replicated measurements. Some tools allow varying attributes of a node other than the fill color. However, browsing through time points is then cumbersome, since two distinct attributes of a node need to be changed (usually manually).

In this paper, we present an application for integrated visualization of gene expression data from time series experiments in gene regulation networks and metabolic networks. Such integration is necessary, since it provides the link between the measurements at the transcriptional level and the observable characteristics of an organism at the functional level. In previous work, we already made a first step towards the integration of microarray data, regulatory networks, and metabolic pathways [22]. The main limitation of this approach was that metabolic pathways are considered as disjoint processes. Most metabolites (i.e. compounds or reactions), however, are shared among several pathways. Therefore, to investigate an organism's metabolism, it is relevant to study all the pathways simultaneously. This requires visualization of the whole metabolic network, which have addressed previously [3]. We now step forward by combining both approaches. Our application can (i) visual-

ize the data from time series experiments in the context of a regulatory network and a metabolic network; (ii) identify and visualize active regulatory subnetworks from the gene expression data; (iii) perform a statistical test to identify and subsequently visualize affected metabolic subnetworks.

The remainder of this paper is organized as follows. In Section 2, we describe visualization of gene regulatory networks and expression data from time series experiments. We also summarize a detection algorithm that identifies active subnetworks in the transcription network based on time series data. Visualization of metabolic networks, and the extraction of active subnetworks is discussed in Section 3. We introduce the new visualization tool in Section 4, and we draw conclusions in Section 5.

2 Genomic level visualization

In previous work, we developed GENeVis, the first application for visualization of networks that supports overlaying time series data with associated statistical data on the nodes [23]. This approach was extended to enable visualization of genome expression and network dynamics in both regulatory networks and metabolic pathways [22]. We will briefly summarize the key aspects of GENeVis below.

2.1 Regulatory network visualization

A gene regulatory network is represented by a graph, in which nodes represent genes, and edges represent interactions between a gene product (a regulator protein) and its target genes. A regulator either inhibits or activates its target, which is represented by decorating the target end of an edge by a bar or an arrow head, respectively. In addition, edges are colored according to interaction type: green for activation, red for inhibition, and grey for unknown interaction. The unknown type is implemented, because biological data are often incomplete. A node is drawn as a rectangular box annotated with the gene name. The layout of the network is computed by a force-directed algorithm.

2.2 Gene expression from time series

DNA microarrays are used to measure the expression levels of thousands of genes simultaneously. In a time series experiment, the gene expressions are measured as a function of time. Gene expression values can either be absolute levels of expression or ratios of expression levels between a test and a reference condition. Ratios are usually log-transformed to obtain comparable scales for ratios above and below 1. To each expression value, a statistical value is associated, which expresses the reliability of the measurement.

As in GENeVis [23], we draw each time point as a colored rectangular glyph. The expression value determines the color, and the reliability value determines the height of the glyph; taller means more reliable. From the perceptual point of view, this mapping is effective, since the stronger perceptual cue of size is used to give reliable data more emphasis than unreliable data. The entire time series for each gene is drawn as a row of glyphs inside its gene box. In addition, or alternatively, the gene expression value can be used as a fill color for the gene box. Which mapping is most effective depends on the analysis task.

To provide the user some insight in the distribution of the expression data, we order the data, and split it into a number of equally-sized subsets, i.e., quantiles. Each subset is assigned a color from the color map. The actual colors in the map depend on the type of expression data: a color map ranging from white to black via yellow and red is used for expression levels, and a bimodal colormap ranging from green to red via black is used for expression ratios.

2.3 Network dynamics visualization

The algorithm to detect active subnetworks is described in detail in [22]. We briefly summarize it in this section. Denote by $L_{g,t}$ and $R_{g,t}$ the expression level and expression ratio (log-transformed) of the gene g at time point t, respectively. The expression level $L_{g,t}$ is *low* if $L_{g,t} < T_m$, *medium* if $T_m \leq L_{g,t} < T_h$, and *high* if $L_{g,t} \geq T_h$, where T_m and T_h are thresholds so that $0 \leq T_m < T_h$. A gene is *differentially expressed* if $|R_{g,t}| \geq T_r$, where $T_r > 0$ is a threshold. The algorithm distinguishes *regulator* and *non-regulator* genes. A regulator gene g is active at time point t if its expression level is high; or if its expression level is medium and $R_{g,t} \geq 0$; or if its expression level is low and $R_{g,t} \geq T_r$. A nonregulator gene is active if it is differentially expressed. After identifying the active genes, the active edges are determined. An edge is marked active if both the source node and target node are active.

The active network is drawn in the context of the complete network. The active nodes and edges are highlighted, and the inactive part of the network is rendered semi-transparent.

3 Metabolic level visualization

3.1 Metabolic network visualization

The main problem when drawing the whole metabolic network is to respect biological conventions for particular topological features (cycle and cascade of reaction) but also to preserve the metabolic pathway information (i.e. reactions and compounds of a pathway have to be drawn in a "small" region). As these pathways often share reactions

and compounds, it is not straightforward to respect the biological conventions, while preserving the metabolic pathway information at the same time.

To overcome this problem, there exist two approaches: with and without node duplication. In the node duplication approach [8, 14], each reaction or compound shared by several pathways is duplicated. The pathways are drawn separately, and they are all shown in a grid-like fashion. By representing each pathway independently, this approach offers only a set of local views (one for each pathway), rather than a global view on the metabolic network as a whole. Moreover, duplication produces drawings where the depicted connectivity does not match the real topology of the network, which may affect correct interpretation.

Because of these drawbacks, the metabolic network visualization in our tool is based on the approach without node duplication presented by Bourqui *et al.* [3]. This algorithm has two main steps: a clustering step and a rendering step. The clustering step computes a set of independent pathways, and it detects particular topological structures, such as cycles and cascades of reactions. Two pathways are considered independent if they do not share any reaction and/or compound. The clustering process can be constrained by the user, who can provide a list of focus pathways. The clustering algorithm then tries to respect the proximity constraint for these pathways, i.e., it will not split up the pathway across multiple clusters if possible. The rendering step draws the clustered graph and the clusters computed in the previous step, while respecting as much as possible the biological drawing conventions. A detailed explanation of the clustering algorithm and rendering procedure is given elsewhere [3].

3.2 Computing the affected subnetwork

The active subnetwork in the regulatory network is computed according to the algorithm described in Section 2.3. This yields a set of genes that are considered active, which we will now use to extract a corresponding affected subnetwork from the whole metabolic network.

The subnetwork is constructed in two steps. The first step identifies the pathways that are affected at each time point. Consider a pathway P containing N^a affected and N^i unaffected reactions (or genes) at a given time point. From N^a and N^i, we compute the probability p that an enrichment of active genes in that pathway can be attributed to chance by Fishers exact test. A pathway is considered affected if $p \leq T_p$, for some threshold T_p. A low probability p means that pathway P is more affected. Per time point, this step yields a set of affected pathways.

The second step constructs a subnetwork by merging all pathways obtained in the first step. The purpose of extracting this subnetwork is to filter out parts of the metabolic network that are not affected at all. This reduces the size of the network considerably, and it simplifies the visualization.

3.3 Visualizing the affected metabolic subnetwork

To visualize the affected metabolic network, we also use a technique analogue to the one described for the active regulatory network visualization. The representation of the metabolic network is somewhat more complex than for the regulatory network, however, since it is a quotient graph containing metanodes (see [3] for more details). A metanode is a node representing a set of vertices of the original network, and it is either a pathway, or a proper subpart of a pathway (all compounds and reactions only belonging to that pathway), or a particular topological structure (cycle or cascade of reactions).

To emphasize affected pathways, we render unaffected pathways at a given time point semi-transparent. A metanode is made transparent if all nodes or metanodes it contains are transparent. As there exists a correspondence between the genes in the transcription network and the enzymes in the metabolic network, we can additionally visualize the gene boxes with expression glyphs in the context of the metabolic network.

4 Integrative visualization

Our application integrates four different types of data: gene annotation, microarray data, a gene regulatory network and a metabolic network, which are all given in separate files. The gene annotation file provides all information about the genes and is represented as a table. In that table, each row corresponds to a single gene and the table columns contain for each gene the locus tag, common gene name, start and end position on the genome and a data field to store additional information, such as the function of a gene. To be able to relate the genes between the data sources, the annotation file should contain appropriate identifiers, which makes this mapping possible. For the microarray data and transcription network, we use locus tags for this purpose. For the metabolic network, we take EcoCyc IDs [12]. The program is flexible, and can use other identifiers as well. The gene regulatory network file contains an adjacency list of genes, which are identified by locus tags. For each edge, the interaction type is given as well. The metabolic network file is in the SBML [6] format, and it contains a.o. the list of compounds and reactions, theirs attributes, and also the decomposition of the network into metabolic pathways. Finally, microarray data are given by two files, one containing expression levels and the other expression ratios. These data are given as tables in which each row corresponds to

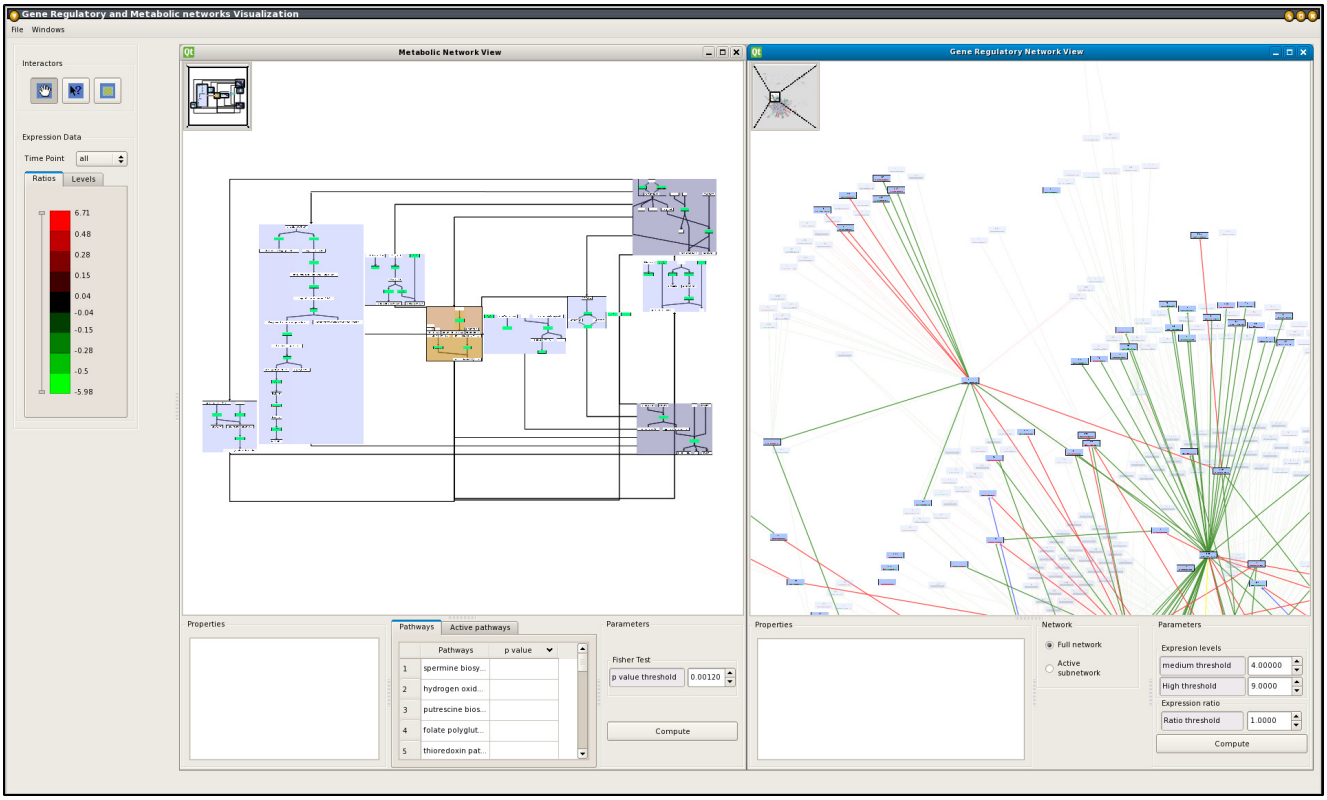

Figure 1. Screenshot of our tool. On the left, some interaction controls are shown. The middle and right panel contain a part of the metabolic network and regulatory network of *E. coli*, respectively. Gene expression data from a 17-points time series experiment are overlaid on the nodes.

a gene, and the columns correspond to the gene expression and associated reliability at all time points.

Figure 1 shows a screenshot of our tool. On the left, some interaction controls are shown. The middle and right panel contain a part of the metabolic network and regulatory network of the bacterium *Escherichia coli K12*, respectively. Gene expression data from a 17-points time series experiment are overlaid on the nodes.

The left panel has three interactor tools shown at the top. From left to right: panning, selection, and rubber band select. In panning mode, the user can move the viewpoint in each network view. Zooming is allowed in either view by simple mouse operations. The selection tool can be used to add single genes to a selection set. Linking and brushing techniques are used to highlight the selected set in both views. The rubber band select tool can be used to select genes in a rectangular area in the transcription network view. This allows creation of subnetworks that can be studied in more detail.

The bottom part of the left panel shows the color legend of the expression data, and allows switching between expression level and expression ratio visualization. A user can also select a time point, which will set the fill color of the gene boxes according to the expression values at that time point. Finally, our tool supports filtering by expression ratio or expression level of genes. This is controlled by a double slider (see Fig. 1 to the left of the color legend). If filtered out, a gene (or a reaction) is considered as unaffected and it will be drawn semi-transparent just as the other unaffected genes (or reactions).

Figure 1 also shows two network visualization widgets: the metabolic network widget on the left and the gene regulatory network widget on the right. Both widgets contain a network view and features dedicated to one of these two types of networks. We will now describe these in more detail.

The gene regulation network widget is shown in Fig. 2. This widget contains four main parts, labelled (A) to (D). The visualization panel (A) offers an overview+detail visualization of the regulatory network. When the user select a gene in panel (A) (or in the metabolic network visualization), its properties are displayed in panel (B). Some genes are never considered affected during any time point, therefore the user may want to remove them from the visualiza-

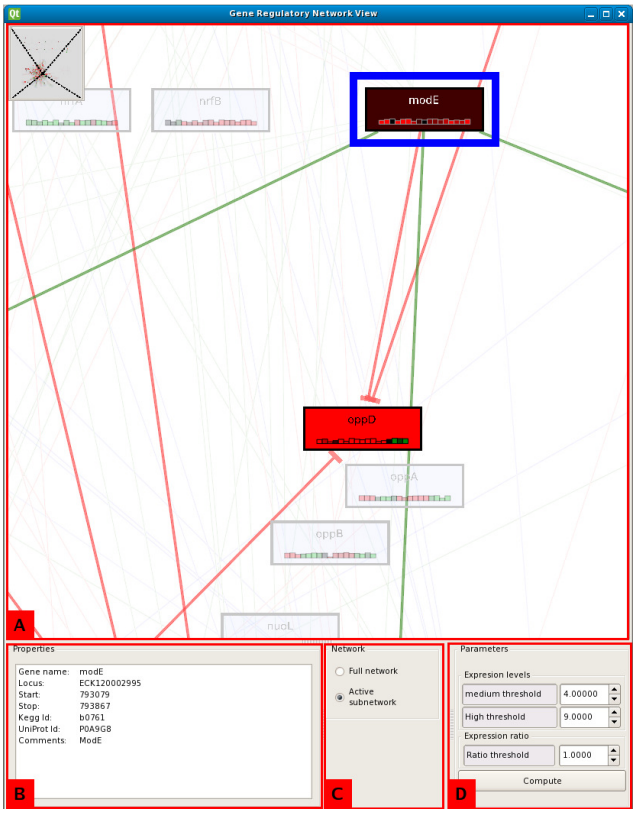

Figure 2. Screenshot of the gene regulatory network widget. (A) The visualization panel. (B) Selected gene properties. (C) Full network visualization or affected network visualization. (D) Affected genes detection parameters.

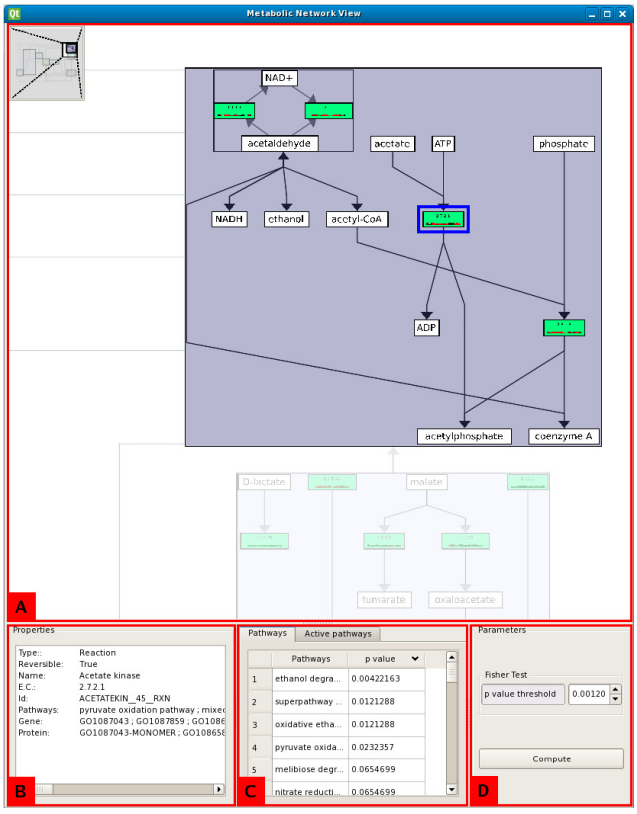

Figure 3. Screenshot of the metabolic network widget. (A) The visualization panel. (B) Selected reaction/compound properties. (C) Lists of unaffected and affected pathways and their corresponding p-value according to Fisher's exact test. (D) Fisher's exact test threshold.

tion. Panel (C) allows to filter out these unaffected genes. Finally, panel (D) allows the user to set up the parameters for the active network detection algorithm.

Figure 3 shows the metabolic network widget. This widget contains four main parts, labelled (A) to (D). Panels (A) and (B) offer the same possibilities as in the gene regulatory widget, i.e. an overview+detail visualization of the network and the properties of the selected compound or reaction. Panel (C) contains two tables, one for the unaffected and one for the affected pathways and their corresponding Fisher's exact test values. Finally, panel (D) allows the user to change the threshold T_p used in Fisher's exact test.

A video demonstration of the application is available at our website[1].

[1] http://www.labri.fr/perso/bourqui/demo_IV09.avi

5 Discussion

We have used a time series dataset from the bacterium *Escherichia coli* [4] to perform an initial validation of our approach. The aim of the validation test was to reproduce findings reported in [4]. The dataset contains gene expression data from 17 time points, during which *E. coli* is grown on a mixture of glucose and lactose. The bacterium grows preferentially on glucose until that energy source is depleted, resulting in growth arrest while the cells adjust to growth on lactose. This shift, called the diauxic shift, takes places at about time point 6. At time point 14, the stationary phase is entered in which the organism stops growing due to the lack of nutrients. During this phase, many processes are shut down by the bacterial cell in order to save energy. The expression levels for this data set range from 0 to 14.97. We set the medium and high expression level thresholds to $T_m = 5.0$ and $T_h = 10.0$, respectively. The

ratio threshold was set to $T_r = 1.5$, and we consider pathways significant at the 5% level, i.e., we set $T_p = 0.05$. We find many pathways related to amino acid degradation and amino acid biosynthesis, which is consistent with the findings of Chang et al. [4]. At the time of the diauxic shift, the genes involved in lactose degradation are upregulated, and we find lactose degradation-related pathways with significance levels around $p = 0.02$. It is beyond the scope of this paper to provide an in-depth analysis of this dataset, but these initial results show that our integrated approach can reproduce results of a traditional approach that involves many manual steps.

In this paper, we have presented an application for visualization of gene expression data from microarray time series experiments in both a gene regulatory network and a metabolic network context. The expression data can be used to study activity at the transcriptional level, and this information can be linked to behavioral characteristics of an organism at the metabolic level. The combination of visual exploration of a time series in multiple contexts and statistical tests for analysis is powerful, and it speeds up the data analysis process. In comparison with other biological network visualization approaches, our application is not dependent on specific databases. Transcription networks and microarray data can be provided in simple flat text files, and any metabolic network provided in SBML format can be loaded. This allows domain experts to load their own data in a straightforward manner.

As part of future work, we will perform an experimental evaluation of the effectiveness and the efficiency of our tool. This will involve domain experts with whom we already have a fruitful collaboration.

We are planning a number of extensions to our current work. To better support data analysis, it is necessary to provide statistical analysis methods, so that a domain expert does not have to rely on visual inspection alone. A commonly-used method involves clustering of expression profiles to find groups of genes that show similar behavior. This helps a biologist in determining functional classes, or it enables inferring gene function for genes with unknown function. It would be useful to visualize such clusters in the context of the transcription and metabolic network.

6 Acknowledgement

This work was partially done under the Expression of Interest project, supported by the VIEW programme of the Netherlands Organisation for Scientific Research (NWO) under research grant no. 643.100.502.

References

[1] M. Baitaluk, M. Sedova, A. Ray, and A. Gupta. BiologicalNetworks: Visualization and analysis tool for systems biology. *Nucleic Acids Res.*, 34:W466–W471, 2006.

[2] M. Becker and I. Rojas. A Graph Layout Algorithm for Drawing Metabolic Pathways. *Bioinformatics*, 17:461–467, 2001.

[3] R. Bourqui, V. Lacroix, L. Cottret, D. Auber, P. Mary, M.-F. Sagot, and F. Jourdan. Metabolic network visualization eliminating node redundance and preserving metabolic pathways. *BMC Systems Biology*, 1(29), 2007.

[4] D. E. Chang, D. J. Smalley, and T. Conway. Gene expression profiling of Escherichia coli growth transitions: an expanded stringent response model. *Molecular Microbiology*, 45(2):289–306, 2002.

[5] U. Dogrusoz, E. Z. Erson, E. Giral, E. Demir, O. Babur, A. Cetintas, and R. Colak. Patikaweb: a web interface for analyzing biological pathways through advanced querying and visualization. *Bioinformatics*, 22(3):374–375, November 2005.

[6] A. Finney and M. Hucka. Systems biology markup language: Level 2 and beyond. *Biochem Soc Trans*, 31(6):1472–1473, 2003.

[7] Z. Hu, D. M. Ng, T. Yamada, C. Chen, S. Kawashima, J. Mellor, B. Linghu, M. Kanehisa, J. M. Stuart, and C. DeLisi. VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. *Nucleic Acids Res.*, 35:W625–W632, 2007.

[8] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33:D428–D432, 2005.

[9] F. Jourdan and G. Melançon. A Tool for Metabolic and Regulatory Pathways Visual Analysis. In *Visualization and Data Analysis, VDA*, pages 46–55. SPIE, jan 2003.

[10] B. H. Junker, C. Klukas, and F. Schreiber. VANTED: A system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 7:109, 2006.

[11] P. D. Karp, C.A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahren, S. Tsoka, N. Darzentas, V. Kunin, and N. Lopez-Bigas. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research*, 19:6083–6089, 2005.

[12] I.M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I.T. Paulsen, M. Peralta-Gil, and P. D. Karp. Ecocyc: A comprehensive database resource for Escherichia coli. *Nucleic Acids Research*, 33:D334–D337, 2005.

[13] B. Mlecnik, M. Scheideler, H. Hackl, J. Hartler, F. Sanchez-Cabo, and Z. Trajanoski. PathwayExplorer: Web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res.*, 33:W633–W637, 2005.

[14] S. Pailey and P. D. Karp. The Pathway Tools cellular overview diagram and Omics Viewer. *Nucleic Acids Research*, 34(13):3771–3778, 2006.

[15] W. Salamonsen, K. Y. Mok, P. Kolatkar, and S. Subbiah. Biojake: a tool for the creation, visualization and manipulation of metabolic pathways. In *Pacific Symposium on Biocomputing*, number 4, pages 392–400, 1999.

[16] P. Saraiya, C. North, and K. Duca. Visualizing biological pathways: Requirements analysis, systems evaluation and research agenda. *Information Visualization*, 4(3):191–205, 2005.

[17] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, 2003.

[18] M. Suderman and M. Hallett. Tools for visually exploring biological networks. *Bioinformatics*, 23(20):2651–2659, 2007.

[19] J. van Helden, L. Wernisch, D. Gilbert, and S. Wodak. Graph-based analysis of metabolic networks. *Ernst Schering Research Foundation Workshop*, 38:245–274, 2002.

[20] K. Wegner and U. Kummer. A new dynamical layout algorithm for complex biochemical reaction networks. *BMC Bioinformatics*, 6(212), 2005.

[21] M. Weniger, J. C. Engelmann, and J. Schultz. Genome expression pathway analysis tool – analysis and visualization of microarray gene expression data under genomic, proteomic and metabolic context. *BMC Bioinformatics*, 8:179, 2007.

[22] M. A. Westenberg, S. A. F. T. van Hijum, O. P. Kuipers, and J. B. T. M. Roerdink. Visualizing genome expression and regulatory network dynamics in genomic and metabolic context. *Computer Graphics Forum*, 27(3):887–894, 2008.

[23] M. A. Westenberg, S. A. F. T. van Hijum, A. T. Lulko, O. P. Kuipers, and J. B. T. M. Roerdink. Interactive visualization of gene regulatory networks with associated gene expression time series data. In L. Linsen, H. Hagen, and B. Hamann, editors, *Visualization in Medicine Life Sciences*, Visualization and Mathematics, pages 293–312. Springer Verlag, Berlin, Germany, 2007.

[24] D. Wolf, C. P. Gray, and A. de Saizieu. Visualizing gene expression in its metabolic context. *Briefings in Bioinformatics*, 1(3):297–304, 2000.